

DOCUMENT RESUME

ED 451 200

TM 032 426

AUTHOR Christmann, Edwin P.; Badgett, John L.
TITLE A Meta-Analytic Comparison between the Assigned Academic Achievement Levels of Students Assessed with Either Traditional or Alternative Assessment Techniques.
PUB DATE 2000-00-00
NOTE 13p.
PUB TYPE Journal Articles (080) -- Reports - Evaluative (142)
JOURNAL CIT Louisiana Education Research Journal; v25 n1 p55-65 1999/2000
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Academic Achievement; Comparative Analysis; *Educational Assessment; Effect Size; Elementary Secondary Education; *Meta Analysis; *Student Evaluation
IDENTIFIERS *Alternative Assessment

ABSTRACT

A meta-analysis was conducted to determine whether there were differences between the assigned academic achievement levels of students who were assessed with traditional methods of assessment and those who were assessed with alternative methods. From the more than 800 studies identified through literature searches, 7 studies, with a total of 5,020 student achievement levels and 15 effect sizes, were selected for the analysis. Findings suggest that efforts to compare the effectiveness of traditional and alternative assessment on academic achievement may be exercises in futility since there was no consensual agreement on the meaning of the term "academic achievement" and there were different connotations for "reliability." However, currently available data suggest a very small, if not trivial, gain for the use of alternative assessment procedures, and given how costly these procedures are, the benefits may not outweigh the costs. (Contains 1 table and 22 references.) (SLD)

A META-ANALYTIC COMPARISON BETWEEN THE ASSIGNED ACADEMIC ACHIEVEMENT LEVELS OF STUDENTS ASSESSED WITH EITHER TRADITIONAL OR ALTERNATIVE ASSESSMENT TECHNIQUES

Edwin P. Christmann, Slippery Rock University
John L. Badgett, Slippery Rock University

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

✓ This document has been reproduced as
received from the person or organization
originating it.

□ Minor changes have been made to
improve reproduction quality.

• Points of view or opinions stated in this
document do not necessarily represent
official OERI position or policy.

PERMISSION TO REPRODUCE AND
DISSEMINATE THIS MATERIAL HAS
BEEN GRANTED BY

E. Christmann

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)

1

TM032426

BEST COPY AVAILABLE

A META-ANALYTIC COMPARISON BETWEEN THE ASSIGNED ACADEMIC ACHIEVEMENT LEVELS OF STUDENTS ASSESSED WITH EITHER TRADITIONAL OR ALTERNATIVE ASSESSMENT TECHNIQUES

**Edwin P. Christmann, Slippery Rock University
John L. Badgett, Slippery Rock University**

INTRODUCTION

Traditional assessments are currently under a siege of criticism from proponents of "authentic" assessment, who insist that the objectivity of paper and pencil tests is incongruent with contemporary classroom instruction (Shepard et al., 1996). Moreover, these critics argue that objective tests prohibit the measurement of higher-order thinking skills (Hasit & DiObilda, 1996). Linn and Gronlund (1995), on the other hand, have provided detailed examples of the effectiveness of objective tests in the measurement of such skills. Further, Brennan and Johnson (1995) also point out that even though "the 'authentic nature' of performance assessments is quite appealing," it should be remembered that "the realism of performance assessments comes at the cost of limitations in the generalizability of results." In agreement, Phillips (1993) criticizes alternative assessments because of their "lack of generalizability from selected tasks to the domain of interest." Hence, it appears that the external validity of alternative, or performance, assessments is in question.

The reliability of performance assessments is also under scrutiny, as seen in Willson's (1991) insistence that this methodology "cannot ignore fundamental psychometric principles of reliability . . .," and Brennan and Johnson (1995) warn that these assessments "raise a host of technical problems that must be faced if annual performance assessments are to yield comparable results from year to year." Nevertheless, Hirsch (1996) reports that advocates of performance assessment proclaim that such assessments are superior to objective tests because they are more informative and motivational, and are also fairer to minorities and nonverbal

students. Concurring, Meisels and Dorfman (1995) assert that minorities -- especially African Americans -- "have not fared well under the domination of multiple choice examinations." And Langer et al. (1990) resound that multiple choice tests measure only recognition and retention "while alternative assessment measures the thinking curriculum." Moreover, Willson (1991) echoes that the perceived weakness of the multiple choice test in assessing higher-order thinking skills has necessitated the development of writing samples "for many state assessments."

The findings of Davis and Felknor (1994), however, disclose that a majority of the students opposes alternative assessments, and only a few feel that these assessments are motivating. Also, Dorfman and Steele (1995) point out that the National Assessment of Education Progress has revealed that the mean differences between blacks and whites on "the extended-response essays" exceed those differences "found on the multiple choice reading assessment." Then, in response to the reported deficiencies of objective examinations in measuring complex thinking skills, Phillips (1993) reminds us that Forsyth (1976) has provided extensive examples of the capacity of objective items to measure higher-order thinking processes; and Phillips (1993) also reminds that Mehrens (1990) has observed that even cognitive psychologists warn against the widespread use of alternative assessment until such theories are documented by extensive research.

Obviously, proponents of both traditional and alternative assessments are insistent that their respective methodologies are more conducive to the academic achievement of America's students. Presently, however, there is no clear consensus favoring either method. Hopefully, this study will provide broader findings that will help resolve the current dilemma.

STATEMENT OF THE PROBLEM

The enhancement of academic achievement in America's schools was the underlying impetus for performing this meta-analysis on all suitable research that has compared traditional with alternative assessments. In compliance with the previously stated purpose of the investigation, this meta-analysis focused on the following research question:

Are there differences between the assigned academic achievement levels of students who were assessed with traditional methods of assessment, and those who were assessed with alternative methods?

METHODOLOGY

The meta-analytic approach used in this study follows the procedure developed by Glass et al. (1981). More specifically, this approach to meta-analysis requires the following: (a) locating studies through unbiased and replicable data searches, (b) selecting studies based on predetermined criteria, (c) describing each study's outcomes and then creating a common scale (effect size), (d) using statistical methods to quantify a specific conclusion from a mixed set of results. Fundamentally, meta-analysis is a quantitative application of empirical deduction that would have been impossible through any other previously known methodology (Gall et al., 1996).

Locating of documents. The studies examined in this research were selected from a computer search of the databases ERIC (1966-March 1999), Dissertation Abstracts (1861-August 1997), and PsychLit (1974-September 1997). These databases were searched with the keywords "alternative assessment," "traditional assessment," "evaluation," and "achievement," which identified over 800 studies to be reviewed for inclusion in the meta-analysis. The included studies met the following predetermined criteria:

1. they were conducted in an educational setting;
2. they included quantitative results in which academic achievement was identified by the author(s) as the dependent variable, and the assessment methodology was the independent variable;
3. they had experimental, quasi-experimental, or correlational research designs;
4. the sample sizes had a combined minimum of 20 students in the experimental and control groups;

5. all academic achievement was reported as interval data;
6. had sufficient statistical data to calculate an effect size.

Coding of the variables. Traditional assessments, for the most part, consist of paper and pencil objective and essay examinations, whereas alternative assessments encompass the evaluation of students, reflective journal writing, group projects, self-assessments, slide shows, oral presentations, writing samples, and so on. Basically, academic achievement is defined as teacher-assigned grades or as student scores on standardized tests. However, all academic achievement included in this meta-analysis was reported in terms of interval data.

Seven of more than 800 relevant publications met the prearranged criteria for inclusion in the meta-analysis, whereas those studies that were rejected did not meet each of the six criteria necessary for incorporation into the study. Generally, those studies not meeting the six prerequisite criteria did not employ statistical analyses. Moreover, if a study employs multiple dependent variables as if they were separate entities, Glass (1981) posits that calculating the multiple effect sizes from such a study is an acceptable procedure for calculating average effect sizes, thus sanctioning the presence of multiple independent comparisons in independent research articles. In compliance with Glass's theoretical methodology of meta-analysis, this study was able to disclose 15 effect sizes from the seven studies examined in its meta-analysis.

ANALYSIS

As previously mentioned, the data were analyzed through a meta-analytic technique, which relies heavily on the calculation of effect sizes for establishing statistical meaning (Wolf, 1986). According to Glass et al. (1981), effect size is the degree to which a phenomenon is present in the population of the study. In meta-analysis (Wolf, 1986), effect size is calculated to determine the presence of a statistical difference between mean standard deviation units (SD_x).

META-ANALYSIS

Seven studies with a total of 5020 student achievement levels and 15 effect sizes generating 15 conclusions met the predetermined criteria for incorporation into the meta-analysis. The individual sample sizes ranged from 25 to 1381, and the mean sample size was 335. Table 1 displays the author(s), date, sample size, standard unweighted mean effect size, and standard error for each of the included studies.

Table 1. Date, Sample Size, and Effect-Sizes

Author(s)	Date	<u>n</u>	<u>ES</u>	<u>SE</u>
Joyce, et al.	1988	286	-0.212	0.006
Laesch, et al.	1987	30	0.517	0.064
Macciomei, N. R.	1995	46	0.031	0.041
Macciomei, N. R.	1995	46	-0.071	0.041
Macciomei, N. R.	1995	46	-0.028	0.041
Saturneli, et al.	1995	1381	1.186	0.001
Seda-Santana, et al.	1988	28	0.561	0.070
Seda-Santana, et al.	1988	25	0.200	0.080
*Shepard, et al.	1996	500	-0.102	0.004
*Shepard, et al.	1996	498	-0.521	0.004
*Shepard, et al.	1996	496	-0.171	0.004
*Shepard, et al.	1996	533	-0.034	0.003
*Shepard, et al.	1996	536	0.101	0.003
*Shepard, et al.	1996	534	-0.171	0.003
Slater, et al.	1995	35	-0.184	0.050

*Estimated effect size calculations are based on pooled, within-school standard deviations.

MEAN EFFECT SIZES

An overall mean effect size was also computed from the 15 calculated effect sizes. The sum of the 15 effect sizes is 0.900, and the mean unweighted effect size was 0.060, with a standard error of 0.030, which is positive, thus indicating that higher achievement levels were attained by those students who were assessed with alternative as opposed to traditional methodology. In addition, an average weighted unbiased estimate of effect size (ES_w) of 0.168 was calculated. However, Cohen (1977) classifies this effect as less than small. Perhaps even more important, the study of Saturnelli et al. (1995), as depicted in Table 1, included the examination of 1381 subjects in arriving at an effect size of 1.186, which was obviously resultant in a positive mean effect size for the total meta-analysis. However, this large positive effect size was offset by the six negative effect sizes disclosed by Shepard et al. (1996).

Nevertheless, Wolf's (1986) interpretation of average unweighted effect in SD units for the comparison between traditional assessment methodology and alternative assessment methodology indicates that the average student exposed to alternative assessment methodology exceeded 52.4% of those students who were exposed to traditional assessments. Moreover, on the basis of an average unweighted and unbiased estimate of effect size, the typical student moved from the 50th percentile to the 52.4th percentile when exposed to alternative assessments. Again, however, any interpretation of these results should be tempered by an awareness of Saturnelli et al.'s (1995) unusually large positive effect size, which was most instrumental in the comparatively higher academic achievement of the groups receiving alternative as opposed to traditional assessments. But the research of Shepard et al. (1996) possibly softened this effect.

DISCUSSION

As mentioned, Cohen's (1977) classification of the mean effect size of 0.060 as less than small is reinforced by Wolf's (1986) indication that: (1) students receiving alternative assessment exceeded 52.4% of those receiving traditional assessment; and (2) a typical student moves from the 50th to the 52.4th percentile when assessed by alternative methodology. How-

ever, given the nature of percentiles, this is a very small, and perhaps trivial, difference. Moreover, these conclusions are somewhat contaminated by the encompassing nomenclature of "academic achievement." More specifically, even teacher-assigned grades that are based on objective test scores differ, as does student performance on separate standardized tests. Then, when teacher evaluations of debatable academic performances such as cooperative learning projects, skits, self-assessments, and reflective journal writings are the basis of student grades, the definition of "academic achievement" becomes further obscured.

Although the effect sizes of the Saturnelli et al. data are positive, and those of Shepard et al. are negative, the two data sets share distinct commonalities. Both included large sample sizes ($n = 1381$; average $n_6 = 516$), both were conducted in urban elementary school settings (New York, Denver), and both contained relatively high minority representations (43%, 41%). However, the two differ with respect to instructional methodology and subject areas in which academic achievement was assessed. Specifically, the data of Saturnelli et al. involved the assessment of science, whereas those of Shepard et al. involved the assessment of pupil performance in reading and mathematics.

Possibly, the extreme difference between the effect sizes of the two data sets lies in the differing teaching methodologies. The positive effect size of the saturnelli et al. data involved the teaching of science per se, probably by traditional methodology. Hence, singular efforts were focused solely on science instruction, rather than on science in conjunction with another academic subject. However, the negative effect sizes of Shepard et al.'s data may well have involved the simultaneous teaching of reading and mathematics, reflecting the methodology of "whole language constructivism."

Perhaps, teaching a basically quantitative subject in conjunction with a qualitative process such as reading could obstruct maximum achievement in both areas, much as a child's simultaneous study of two different languages could restrict her optimal learning of each language, as opposed to studying the two separately. Then, possibly compounding the situation, minority students are reported to score comparatively lower on alternative assessments (Davis et al., 1994). However, it must be empha-

sized that these explanations are simply conjectural, and obviously subject to further research.

CONCLUSIONS AND RECOMMENDATIONS FOR FURTHER RESEARCH

It may be that efforts to compare the effectiveness of traditional and alternative assessment on academic achievement are "exercises in futility." Initially, there is no consensual agreement between proponents of the two assessment methodologies on the term "academic achievement." Furthermore, since "reliability" has differing connotations for quantitative and qualitative methodologists, a legitimate comparison becomes even more questionable.

Yet, further comparisons are definitely needed, and it is the opinion here that such comparisons are indeed both possible and necessary. However, given the currently available data, only very small (if not trivial) gains, at best, can be attributed to the use of alternate assessment procedures; and given how labor-intensive these procedures are, the benefits do not necessarily outweigh the costs. Nevertheless, it seems that if recipients of alternative assessments were allowed to engage in traditional assessment procedures for a one-to-two-week period, then equivalent grounds for a comparison on the basis of objective measurement could be established. Granted, it is acknowledged that since all school performance is not academic, objective measurement is not always possible. Nevertheless, it would appear that such a proposal could provide for an authentic academic, if not affective, comparison between traditional and alternative assessment. In any event, this procedure could be conducive to the further enhancement of assessment in contemporary American schools.

REFERENCES

(References marked with an asterisk indicate studies included in the meta-analysis.)

- Brennan, R. & Johnson, E. (1995). Generalizability of performance assessments. *Educational Measurement: Issues and Practice*, 15(3), 9-12.

- Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.
- Davis, A. & Felknor, C. (1994). The demise of performance-based graduation in Littleton. *Educational Leadership*, 51(5), 64-65.
- Forsyth, R. (1976). *Describing what Johnny can do*. Iowa City, Iowa: Iowa Testing Programs.
- Gall, M., Borg, W., & Gall, J. (1996). *Educational research: An introduction*. White Plains, NY: Longman.
- Glass, G., McGaw, B., & Smith, M. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.
- Hasit, C. & DiOblida, N. (1996). Portfolio assessment in a college developmental reading class. *Journal of Developmental Education*, 19(3), 26-31.
- Hirsch, E. D. (1996). *The schools we need: Why we don't have them*. New York: Doubleday.
- *Joyce, B. G. & Wolking, W. D. (1988). Curriculum-based assessment: An alternative approach for screening young gifted children in rural areas. *Rural Special Education Quarterly*, 8(4), 9-14.
- *Laesch, K. B. & Van Kleeck, A. (1987). The cloze test as an alternative measure of language proficiency of children considered for exit from bilingual education programs. *Language Learning*, 37(2), 172-189.
- Langer, J. A., Applebee, A. N., Mullis, I., & Foertsch, M. (1990). *Learning to read in our nation's schools: Instructional and achievement in 1988 at grades 4, 8, and 12*. Princeton, NJ: Educational Testing Service.
- Linn, R. & Gronlund, N. (1995). *Measurement and assessment in teaching*. Upper Saddle River, NJ: Prentice Hall.

- *Macciomei, N. R. (1995). The effect of portfolio assessment on academic achievement and intrinsic motivation for students with specific learning disabilities. (Doctoral Dissertation, University of South Carolina.) *Dissertation Abstracts International*, 56, 12A.
- Mehrens, W. (1990). New tests based on performance raise questions. *Education Week*, 19(2), 12.
- Meisels, S. J., Dorfman, A., & Steele, D. (1995). Equity and excellence in group administered and performance based assessments. In M. T. Nettles & A. L. Nettles (Eds.), *Equity and Excellence in Educational Testing and Assessment* (pp. 243-261). Boston: Kluwer.
- Phillips, S. E. (1993). Legal issues in performance assessment. *West's Education law Reporter*, 79(3), 709-738.
- *Saturnelli, A. M. & Repa, J. T. (1995, April). *Alternative forms of assessment in elementary science: The interactive effects of reading, race, economic level and the elementary science specialist on hands-on and multiple choice assessment of science process skills*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.
- *Seda-Santana, I. (1988). Retellings and semi-structured interviews for assessing reading comprehension of standardized test passages and the Illinois inventory of educational progress passages as compared to scores on multiple choice test items. (Doctoral Dissertation, university of Illinois at Urbana-Champaign). *Dissertation Abstracts International*, 50, 02A.
- *Shepard, L. A., Flexer, R. J., Hiebert, E. H., Marion, S. F., Mayfield, V. & Weston, T. J. (1996). Effects of introducing classroom performance assessments on student learning. *Educational Measurements: Issues and Practice*, 15(3), 7-18.
- *Slater, T. F., Samson, S. L., & Ryan, J. M. (1995, April). *A qualitative and quantitative comparison of the impact of portfolio assessment procedures versus traditional assessment in a college physics course*.

Paper presented at the annual meeting of the National Association for Research in Science Teaching, San Francisco, CA.

Willson, V. L. (1991). Performance assessment, psychometric theory and cognitive learning theory: Ships crossing in the night. *Contemporary Education*, 62(4), 250-254.

Wolf, F. (1986). *Meta-analysis: Quantitative methods for research synthesis*. Newbury Park, London: Sage.